Supporting information for Decoding asymptomatic COVID-19 infection and transmission

Rui Wang[1], Jiahui Chen[1], Yuta Hozumi [1], Changchuan Yin[2], and Guo-Wei Wei[1,3,4*]
[1] Department of Mathematics,
Michigan State University, MI 48824, USA.
[2] Department of Mathematics, Statistics, and Computer Science,
University of Illinois at Chicago, Chicago, IL 60607, USA
[3] Department of Electrical and Computer Engineering,
Michigan State University, MI 48824, USA.
[4] Department of Biochemistry and Molecular Biology,
Michigan State University, MI 48824, USA.

November 3, 2020

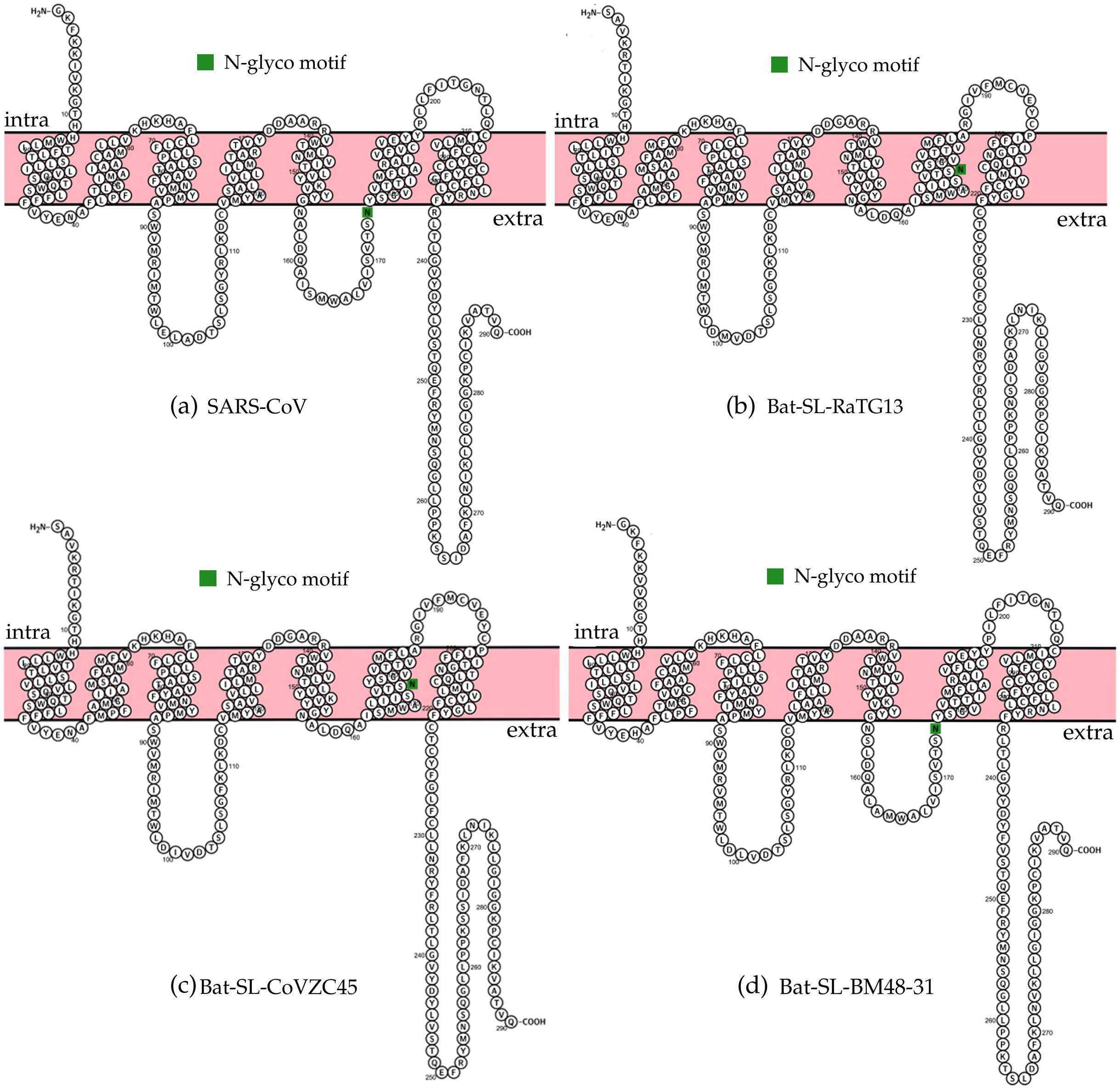
Contents

[*] Corresponding author. E-mail: weig@msu.edu

S1 Additional Analysis

S1.1 Visualization of four coronaviral NSP6s in membrane

Figure S1 is the visualization of four NSP6 proteoforms of SARS-CoV, Bat-SL-RaTG, Bat-SL-CoVZC45, and Bat-SL-BM48-31. These proteoforms are consistent with that of SARS-CoV-2 described in the main paper, indicating their similar function of regulating cell autophagy.



(a) SARS-CoV

(b) Bat-SL-RaTG13

(c) Bat-SL-CoVZC45

(d) Bat-SL-BM48-31

Figure S1: The visualization of NSP6 proteoforms for a few species close to SARS-CoV-2. (a) SARS-CoV, (b) Bat-SL-RaTG13, (c) Bat-SL-CoVZC45, and (d) Bat-SL-BM48-31.

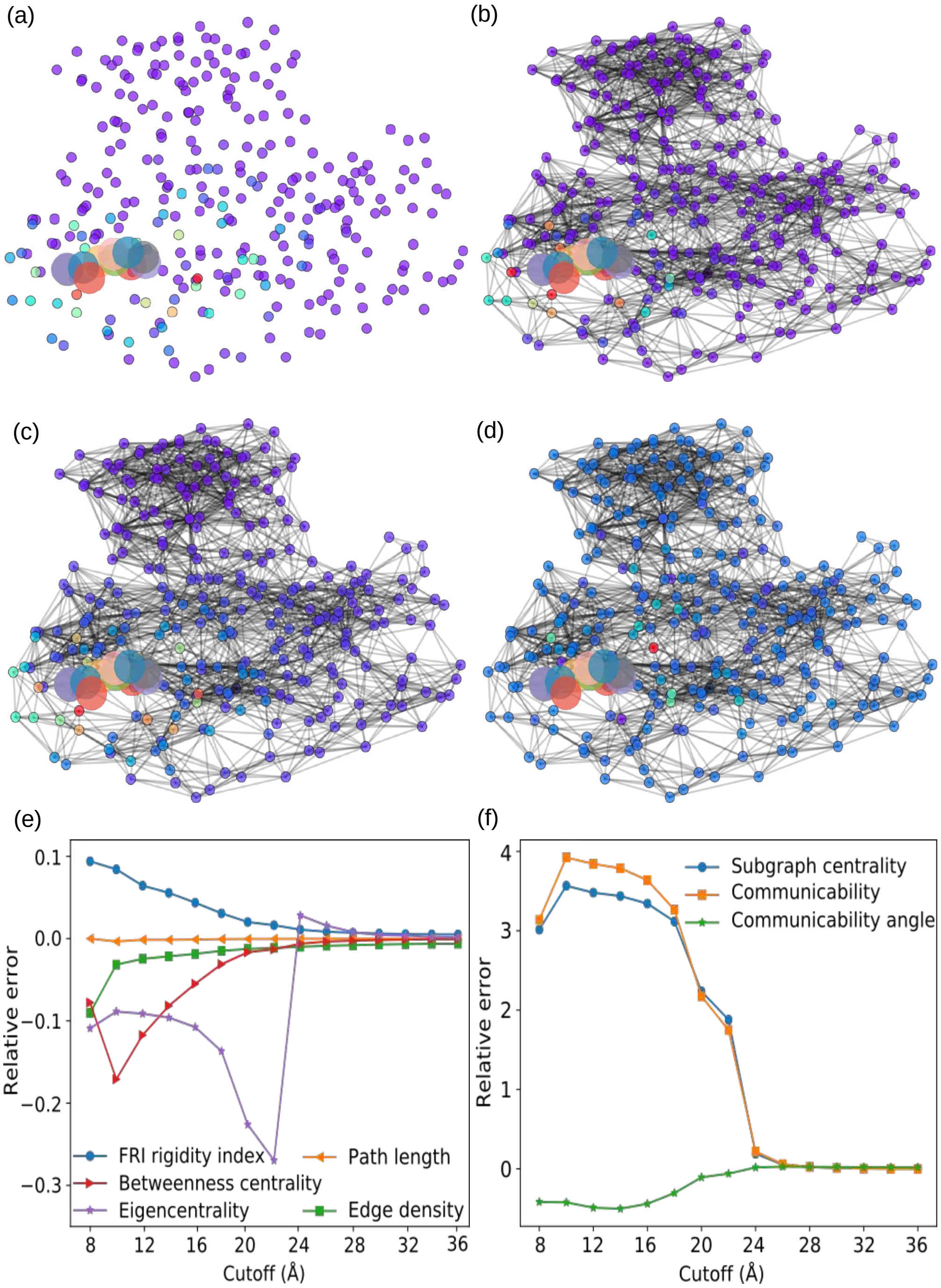


Figure S2: Network analysis of the SARS-CoV-2 NSP6 L37F mutation. The networks consist of heavy atoms of mutation site 37 and C_α atoms of SARS-CoV-2 NSP6. The differences of descriptors between the network with wild type, leucine, and the network with mutant type, phenylalanie, are displayed. (a) FRI rigidity index differences; (b) eigencentrality differences; (c) subgraph centrality differences; and (d) betweenness centrality differences. In (e) and (f), relative changes versus cutoff distances to the mutation site are studied. (e) relative changes of FRI rigidity index, path length, edge density, betweenness centrality, and eigencentrality; and (f) relative changes of subgraph centrality, communicability, and communicability angle.

S1.2 Network analysis of the SARS-CoV-2 NSP3 L37F mutation

Figure S2 shows the network analysis of the SARS-CoV-2 L37F mutation. The networks are constructed by heavy atoms at the mutation site and at the C_α atoms of SARS-CoV-2 NSP6. Figures S2 (a), (b), (c), and (d) present the mutation-induced differences of FRI rigidity index differences, eigencentrality differences, subgraph centrality differences, and betweenness, respectively. Although these descriptors do not change much globally, their local changes around the mutation site reveal the L37F mutation-induced stress around the mutation site.

Figures S2 (e) and (f) show the dependence of relative changes over cutoff distances. First of all, every descriptor converges as the cutoff distance increase. Relative changes do not fluctuate for cutoff distance greater than 24 Å. The relative change of FRI rigidity index monotonically decreases in absolute value as the cutoff increases. Similarly, the edge density has a monotonically decreasing pattern. The two networks are close in the path length descriptor. Interestingly, betweenness centrality descriptor has the largest difference at 10 Å cutoff distance, while the eigencentrality descriptor has the largest difference at 22 Å cutoff distance. For betweenness centrality, it shows that the mutation has the largest impact in the network of C_α within 10 Å to any atoms of the target residue. In Figure S2 (f), three descriptors show a similar pattern — they increase in absolute value first and then decrease eventually as the cutoff increases. Overall, the relative change plots indicate that the mutation happening at L37 has the largest impact on the C_α within 10 Å to any atoms of the target residue.

S2 Material and Methods

S2.1 Data collection and pre-processing

On January 5, 2020, the complete genome sequence of SARS-CoV-2 was first released on GenBank (Access number: NC_045512.2) by Zhang's group at Fudan University [21]. Since then, there has been a rapid accumulation of SARS-CoV-2 genome sequences. In this work, 20,656 complete genome sequences with high coverage of SARS-CoV-2 strains from the infected individuals in the world were downloaded from the GISAID database [17] (https://www.gisaid.org/) as of June 19, 2020. All the records in GISAID without the exact submission date were not taken into considerations. To rearrange the 20,656 complete genome sequences according to the reference SARS-CoV-2 genome, multiple sequence alignment (MSA) is carried out by using Clustal Omega [18] with default parameters.

The amino acid sequence of NSP6 is downloaded from GenBank [1]. The three-dimensional (3D) structure of nonstructure protein 6 (NSP6) in this work is generated by I-TASSER model [23]. The 3D structure graph is created by using PyMOL [7].

S2.2 Single nucleotide polymorphism genotying

Single nucleotide polymorphism (SNP) genotyping measures the genetic variations between different members of a species. Establishing the SNP genotyping method to the investigation of the genotype changes during the transmission and evolution of SARS-CoV-2 is of great importance [19, 24]. By analyzing the rearranged genome sequences, SNP profiles, which record all of the SNP positions in teams of the nucleotide changes and their corresponding positions, can be constructed. The SNP profiles of a given SARS-CoV-2 genome isolated from a COVID-19 patient capture all the differences from a complete reference genome sequence and can be considered as the genotype of the individual SARS-CoV-2.

S2.3 Topology-based prediction of protein folding stability changes upon mutation

In this work, the prediction of NSP6 folding energy changes upon mutation is computed by using the topology based mutation predictor (TML-MP) (https://weilab.math.msu.edu/TML/TML-MP/) which is briefly reviewed as following and its detail can be found in the literature [3]. TML-MP applies element specific persistent homology, which reveals essential biological information [5,8]. The method employs the element-specific persistent homology [4] and other biological and chemical properties as machine learning features to train gradient boosted regression tree (GBRT) models. The dataset includes 2648 mutations instances in 131 proteins provided by Dehouck et al [6]. The error analysis based on the dataset is given as Pearson correlations coefficient (R_p) of 0.79 and root mean square error (RMSE) of 0.91 kcal/mol from previous work [3].

As the persistent homology widely applied in a variety of practical feature generation problems, it is also successful in the implementation of predictions of protein folding energy changes upon mutation. The key idea in TML-MP is to use the element-specific persistent homology (ESPH) which distinguishes different element types of biomolecules when building persistent homology barcodes. For instance, commonly occurring protein element types include C, N, O, S, and H. However, hydrogen atoms are often absent from PDB data and sulfur atoms are too few in most proteins to be statistically important. Thus, C, N, and O elements are considered on the ESPH in protein characterization. As for persistent homology, barcodes generated based on ESPH provide a topological representation of molecular interactions. Features are extracted from the different dimensions of persistent homology barcodes by dividing barcodes into several equally spaced bins, which is called binned barcode representation. The auxiliary features such as geometry, electrostatics, amino acid types composition, and amino acid sequence are also included for machine learning training.

The element specific persistent homology is built by adopting the distance function $DI(A_i, A_j)$ describing the distance between two atoms A_i and A_j defined as

$$DI(A_i, A_j) = \begin{cases} \infty, \text{if Loc}(A_i) = \text{Loc}(A_j), \\ DE(A_i, A_j), \text{otherwise}, \end{cases} \tag{S1}$$

where Loc$(\cdot)$ denotes the location of an atom which is either in a mutant site or in the rest of the protein and $DE(\cdot,\cdot)$ is the Euclidean distance between the two atoms. Then, the persistent homology uses simplicial complexes with a specific rule such as Vietoris-Rips complex, Cech complex, or alpha complex. Vietoris-Rips complex (VC) is used for characterizing first-order interaction where alpha complex (AC) is used for characterizing higher-order patterns. Using ESPH to characterize interactions of different kinds, we construct persistent homology barcodes on the atom sets by selecting one certain type of atoms in mutation site and one other certain type of atoms in the rest of the protein. The set of barcodes from one persistent homology computation as $V^{p,d,b}_{\gamma,\alpha,\beta}$ where

- $p \in \{\text{VC}, \text{AC}\}$ is the complex rule,
- $d \in \{DI, DE\}$ is the distance function,
- $b \in \{0, 1, 2\}$ is the topological dimensions,
- $\gamma \in \{\text{M}, \text{W}\}$ is the protein of mutant type or wild type,
- $\alpha \in$ {C,N,O} is the element type selected in proteins except in the mutation site,
- $\beta \in$ {C,N,O} is the element type selected in the mutation residue.

These barcodes are capable of reflecting the molecular mechanism of protein stability. Features are extracted from the groups of persistent homology barcodes. For 18 groups of Betti-0 ESPH barcode such that 9 groups are from the mutant type and 9 groups are from the wild type, one can specify a fixed length interval to divide the ESPH barcodes into a number of equally spaced bins. For example, a length set, $\{[0, 0.5], (0.5, 1], ..., (5.5, 6]\text{Å}\}$ would turn the 18 groups of Betti-0 ESPH barcode into 18*12 features with dimension of the number of atoms. The death and birth of bars are counted in each bin resulting in features. Therefore, this representation enables us to precisely characterize hydrogen bonds, van der Waals, electrostatic, hydrophilic, and hydrophobic interactions. For the higher-order Betti numbers, the emphasis is given on patterns of both short and long-distance scales. Statistics feature are computed for each group of barcodes for Betti-1 and Betti-2, which are sum, max, and the average of bar length, and max and min of birth and death values. Overall, 12*18 features are generated by Betti-0 on VC, and 7*2*18 features are generated by Betti-1 and Betti-2 on AC.

In TML-MP, gradient boosted regression trees (GBRTs) [14] are employed to train the dataset according to the size of the training dataset, absence of model overfitting, non-normalization of features, and ability of nonlinear properties. The GBRT method produces a prediction model as an ensemble method which is a class of machine learning algorithms. It builds a popular module for regression and classification problems from weak learners. By the assumption that the individual learners are likely to make different mistakes, the method uses a summation of the weak learners to eliminate the overall error. Furthermore, a decision tree is added to the ensemble depending on the current prediction error on the training dataset. Therefore, this method is relatively robust against hyperparameter tuning and overfitting, especially for data with a moderate number of features. The GBRT is shown for its robustness against overfitting, good performance for moderately small data sizes, and model interpretability. The current work uses the package provided by scikit-learn (v 0.23.0) [16].

S2.4 Graph network models

The graph network descriptors are briefly presented which are applied in this work. Graph networks can mimic interactions between pairs of units in molecules. The quantify features of the networks can reveal the biological and chemical properties measured by comparing descriptors on different networks. To detect the single residue impact following mutation, the network consists of a set $S(r)$ of C$_\alpha$ atoms from every residue of protein structure except the target mutation residue where r is the cutoff distance such that a C$_\alpha$ atom is included if it is within r Å to any atom of the target mutation. The total atom set $T(r)$ is defined as the atoms (C, N, and O) of the target residue and C$_\alpha$ atoms of $S(r)$. Moreover, two vertices are connected in the network if their distance is less than 8 Å. Thus the adjacency matrix A can be defined as well where A is a matrix containing 0 and 1 such that $A(i,j)=0$ if i-th and j-th are disconnected and $A(i,j)=1$ if i-th and j-th are connected.

S2.4.1 FRI rigidity index

FRI rigidity index was introduced to reflect the flexibility between atoms for molecular interaction prediction [15, 22]. The single residue molecular rigidity index measures its influence on the set $S(r)$ which is given as

$$R_\eta = \sum_{i=1}^{N_S}\sum_{j=1}^{N} e^{-\left(\frac{\|\mathbf{r}_i-\mathbf{r}_j\|}{\eta}\right)^2}, \tag{S2}$$

where N_S is the number of C$_\alpha$ atoms of the set $S(r)$ and N is the number of atoms in total atom set $T(r)$.

S2.4.2 Edge density

Edge density is defined based on the adjacency matrix of the total atom set $T(r)$ such as

$$d = \frac{1}{N_S}\sum_{i=1, i\notin I_T}^{N}\sum_{j=1}^{N} A(i,j), \tag{S3}$$

where I_T is the index set of the mutation residue.

S2.4.3 Average path length

Average path length measures the separation between two vertices of the whole network, which can be used to study infectious diseases spreading in the networks [20]. The average path length for the single mutation system of biomolecular is defined as

$$\langle L\rangle = \frac{1}{2N_S(N-1)} \sum_{i=1, i\notin I_T}^{N} \sum_{j=1}^{N} d(i,j), \tag{S4}$$

where $d(i,j)$ is the shortest path length between vertices v_i and v_j.

S2.4.4 Average betweenness centrality

Average betweenness centrality shows communications in a network [13]. The average betweenness centrality is given as

$$\langle C_b\rangle = \frac{1}{N_S} \sum_{k=1, k\notin I_T}^{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{g_{ij}^k}{g_{ij}}, \tag{S5}$$

where g_{ij}^k is defined as the number of the shortest path between vertices v_i and v_j that passes v_k, and g_{ij} is the number of shortest paths between v_i and v_j.

S2.4.5 Average egeincentrality

Average egeincentrality is the average of elements of the eigenvector V_{max}, which is corresponding to the largest eigenvalues of the adjacency matrix [2] such as

$$\langle C_e\rangle = \frac{1}{N_S} \sum_{i=1, i\notin I_T}^{N} e_i. \tag{S6}$$

S2.4.6 Average subgraph centrality

Average subgraph centrality is built on the exponential of the adjacency matrix, $E = e^A$. The subgraph centrality is the summation of weighted closed walks of all lengths starting and ending at the same node [9, 12]. Thus the average subgraph centrality reveal the average of participating rate of each vertex in all subgraph, which is given as

$$\langle C_s\rangle = \frac{1}{N_S} \sum_{i=1, i\notin I_T}^{N} E(i,i). \tag{S7}$$

S2.4.7 Average communicability

Average communicability is defined in a similar way as the subgraph centrality on the exponential of the adjacency matrix [9–11], which is

$$\langle M\rangle = \frac{1}{N_S(N-1)} \sum_{i=1, i\notin I_T}^{N} \sum_{j=1}^{N} E(i,j), \tag{S8}$$

S2.4.8 Average communicability angle

Average communicability angle is given by [11]

$$\langle \Theta\rangle = \frac{1}{N_S(N-1)} \sum_{i=1, i\notin I_T}^{N} \sum_{j=1}^{N} \theta(i,j), \tag{S9}$$

where $\theta(i,j) = \cos^{-1}\left(\frac{E(i,j)}{\sqrt{E(i,i),E(j,j)}}\right)$.

S3 Supplementary Tables

Total 16 spreadsheets are merged in the Supporting_Tables.xlsx.

Table S1: S1_snpRecords_10192020_All: The GISAID IDs in the world (Up to October 19, 2020).

Table S2: S2_snpRecords_101912020_11083G>T: The GISAID IDs in the world with 11083G>T-(L37F). (Up to October 19, 2020).

Table S3: S3_Transmission_11083G>T: The date that the 11083G>T-(L37F) was first collected in 75 countries (Up to October 19, 2020).

Table S4: S4_11083G>T_ratio: The ratio of 11083G>T in 75 countries (Up to October 19, 2020).

Table S5: S5_CountryDate_11083G>T_14days: The 11083G>T counts of 75 countries in 14-days period (Up to October 19, 2020).

Table S6: S6_StateDate_11083G>T_14days: The 11083G>T counts of 35 states in the United States in 14-days period (Up to October 19, 2020).

Table S7: Acknowledgement table provided by GISAID in Jan 2020.

Table S8: Acknowledgement table provided by GISAID in Feb 2020.

Table S9: Acknowledgement table provided by GISAID in March 2020.

Table S10: Acknowledgement table provided by GISAID in April 2020.

Table S11: Acknowledgement table provided by GISAID in May 2020.

Table S12: Acknowledgement table provided by GISAID in June 2020.

Table S13: Acknowledgement table provided by GISAID in July 2020.

Table S14: Acknowledgement table provided by GISAID in August 2020.

Table S15: Acknowledgement table provided by GISAID in September 2020.

Table S16: Acknowledgement table provided by GISAID in October 2020.

S4 Supplementary Figures

Total 105 figures are compressed in the Supporting_Figures.zip. 73 figures are the bar plots in 73 countries which show the number of counts having L37F mutation counts and the other counts without L37F mutation. Each bar width covers a 10-day period. The other 32 figures are the bar plots of 32 states in the United States.

References

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic acids research*, 37(suppl_1):D26–D31, 2009.

[2] P. Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

[3] Z. Cang and G.-W. Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33(22):3549–3557, 2017.

[4] Z. Cang and G.-W. Wei. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International journal for numerical methods in biomedical engineering*, 34(2):e2914, 2018.

[5] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[6] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, and M. Rooman. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, 25(19):2537–2543, 2009.

[7] W. L. DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1):82–92, 2002.

[8] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pages 454–463. IEEE, 2000.

[9] E. Estrada. Topological analysis of SARS-CoV-2 main protease. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):061102, 2020.

[10] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77(3):036111, 2008.

[11] E. Estrada and N. Hatano. Communicability angle and the spatial efficiency of networks. *SIAM Review*, 58(4):692–715, 2016.

[12] E. Estrada and J. A. Rodriguez-Velazquez. Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103, 2005.

[13] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[14] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[15] D. D. Nguyen, K. Xia, and G.-W. Wei. Generalized flexibility-rigidity index. *The Journal of chemical physics*, 144(23):234106, 2016.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[17] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13), 2017.

[18] F. Sievers and D. G. Higgins. Clustal omega. *Current Protocols in Bioinformatics*, 48(1):3–13, 2014.

[19] R. Wang, Y. Hozumi, C. Yin, and G.-W. Wei. Decoding SARS-CoV-2 transmission, evolution, and ramification on COVID-19 diagnosis, vaccine, and medicine. *Journal of Chemical Information and Modeling*, page https://doi.org/10.1021/acs.jcim.0c00501, 2020.

[20] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[21] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

[22] K. Xia, K. Opron, and G.-W. Wei. Multiscale multiphysics and multidomain models—Flexibility and rigidity. *The Journal of chemical physics*, 139(19):11B614_1, 2013.

[23] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The i-tasser suite: protein structure and function prediction. *Nature methods*, 12(1):7–8, 2015.

[24] C. Yin. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*, 2020.